Skeletal Radiology
https://doi.org/10.1007/s00256-022-04160-z

SCIENTIFIC ARTICLE

A deep learning algorithm for detecting lytic bone lesions of multiple myeloma on CT

Shahriar Faghani[1] · Francis I. Baffour[2] · Michael D. Ringler[2] · Matthew Hamilton-Cave[3] · Pouria Rouzrokh[1] · Mana Moassefi[1] · Bardia Khosravi[1] · Bradley J. Erickson[1]

Received: 29 June 2022 / Revised: 10 August 2022 / Accepted: 11 August 2022

Abstract

Background Whole-body low-dose CT is the recommended initial imaging modality to evaluate bone destruction as a result of multiple myeloma. Accurate interpretation of these scans to detect small lytic bone lesions is time intensive. A functional deep learning) algorithm to detect lytic lesions on CTs could improve the value of these CTs for myeloma imaging. Our objectives were to develop a DL algorithm and determine its performance at detecting lytic lesions of multiple myeloma.

Methods Axial slices (2-mm section thickness) from whole-body low-dose CT scans of subjects with biochemically confirmed plasma cell dyscrasias were included in the study. Data were split into train and test sets at the patient level targeting a 90%/10% split. Two musculoskeletal radiologists annotated lytic lesions on the images with bounding boxes. Subsequently, we developed a two-step deep learning model comprising bone segmentation followed by lesion detection. Unet and "You Look Only Once" (YOLO) models were used as bone segmentation and lesion detection algorithms, respectively. Diagnostic performance was determined using the area under the receiver operating characteristic curve (AUROC).

Results Forty whole-body low-dose CTs from 40 subjects yielded 2193 image slices. A total of 5640 lytic lesions were annotated. The two-step model achieved a sensitivity of 91.6% and a specificity of 84.6%. Lesion detection AUROC was 90.4%.

Conclusion We developed a deep learning model that detects lytic bone lesions of multiple myeloma on whole-body low-dose CTs with high performance. External validation is required prior to widespread adoption in clinical practice.

Keywords Multiple myeloma · Deep learning · Whole-body low-dose CT · Bone segmentation · Lesion detection

Abbreviations

IMWG	International Myeloma Working Group
CNN	Convolutional neural network
DL	Deep learning
DICOM	Digital Imaging and Communications in Medicine
PNG	Portable Network Graphics
YOLO	You Only Look Once
AUROC	Area under the receiver operating characteristic curve
DSC	Dice similarity coefficient
mAP	Mean average precision
SD	Standard deviation

Introduction

Bone destruction, present in approximately 80% of patients with newly diagnosed multiple myeloma, is a hallmark of the disease [1]. On CT, it manifests as lytic lesions in the axial and proximal appendicular skeleton. The extent of lytic bone disease reflects the grade of tumor burden [1], and therefore, in recent International Myeloma Working Group (IMWG) recommendations for the diagnosis of multiple myeloma, the radiographic presence of osteolytic lesions is considered a myeloma defining event [2].

Imaging protocols for whole-body low-dose CT scans typically extend from the vertex of the skull to the knees and include the proximal humeri. The IMWG criteria also recommend axial image reconstructions at 2 or 3 mm

Bradley J. Erickson is the senior author.

✉ Francis I. Baffour
baffour.francis@mayo.edu

[1] Artificial Intelligence Laboratory, Department of Radiology, Mayo Clinic, 200 1st St. SW, Rochester, MN 55905, USA

[2] Division of Musculoskeletal Radiology, Department of Radiology, Mayo Clinic, Rochester, MN, USA

[3] Mayo Clinic Alix School of Medicine, Rochester, MN, USA

Published online: 18 August 2022

section thickness [3]. In our practice, the image dataset of axial reconstructions at 2-mm section thickness results in approximately 1500 image slices. Lytic lesions characteristic of multiple myeloma can be as small as 5 mm [2]. As a result, a clinically significant lytic lesion may potentially be detected on fewer than 5 image slices. The clinical relevance of a single lesion is not negligible. For instance, in a patient with smoldering multiple myeloma or monoclonal gammopathy of undetermined significance, such a lesion upstages the patient to active multiple myeloma that may need treatment. Multiple independent interpretations of these CT scans can improve the rate of lesion detection [4, 5] but are costly.

Deep learning (DL) algorithms can be utilized for object detection and semantic segmentation in medical imaging [6–11]. Convolutional neural networks (CNNs) have been investigated for the pathologic diagnosis of multiple myeloma [12]. Radiologic diagnostic evaluations have also been investigated using magnetic resonance imaging [13] and positron emission tomography/CT [14]. To the best of our knowledge, no study has investigated the role of a CNN model for multiple myeloma lesion detection using CT images [15].

Low-dose whole-body CT is the recommended initial imaging modality for the detection of bone involvement in patients with a new diagnosis of multiple myeloma, and also when disease relapse is suspected. A functional DL algorithm to detect lytic lesions on CTs could improve the value of these CTs for myeloma imaging. Our objective was to develop a DL algorithm and determine its performance at detecting lytic lesions of multiple myeloma.

Materials and methods

Institutional review board approval was obtained for this study, and informed consent was waived. An end-to-end, two-step approach utilizing two DL-based algorithms was preferred for this task [16–18]. Firstly, a bone segmentation algorithm selects the osseous skeleton and eliminates other tissues to provide a specific region of interest with less background noise. Secondly, an object detection algorithm identifies lytic lesions of myeloma in the segmented osseous skeleton.

Subjects

The image datasets of whole-body low-dose CTs between October 2019 and March 2020 from consecutive patients with biochemically confirmed plasma cell dyscrasias were included in this study. The datasets were assigned a random number (Python random module, version 3.8).

Whole-body low-dose CT images

Whole-body low-dose CT at our institution is a single scan from above-raised elbows through the knees using one of three CT scanners (Siemens Edge, Edge + or Force, Forcheim, Germany). CT acquisition parameters were as follows: tube potential of 120 kV, 128 × 0.6 or 192 × 0.6 mm collimation, rotation time of 0.5 s with a pitch of 1, and slice increment of 1 mm. Two image series are reconstructed at 2-mm section thickness and 1-mm skip using a 512 × 512 matrix: a smooth kernel (Br44) series for bone marrow and soft tissue evaluation and a sharp kernel (Br64) series to evaluate the osseous structures. Only the Br64 image series was used for this study.

Dataset splitting

The data were split into train and test sets targeting 90% and 10% splits. Separate train datasets were used for segmentation and lesion detection. Seventy-five percent of image data was used for the lesion detection algorithm development and 25% for the bone segmentation algorithm development since bone segmentation is a simpler task (Fig. 1). To have more robust results, we performed fivefold cross-validation for the lesion detection. Due to the small size of the segmentation

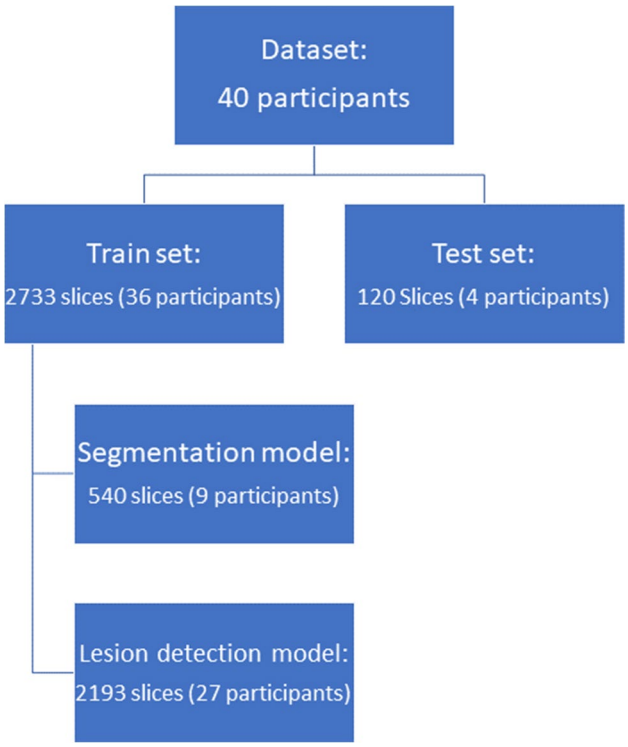

Fig. 1 Flowchart of the data selection process and the composition of the final datasets

dataset, one cross-validation for the segmentation model was excluded [19]. Non-consecutive random image slices were selected from various anatomic regions for the training, validation, and test steps.

Data preprocessing and annotation for skeleton segmentation

Images selected for segmentation were converted from Digital Imaging and Communications in Medicine (DICOM) to Neuroimaging Informatics Technology Initiative, a segmentation-compatible format. Subsequently, the data were divided into anatomical segments from the vertex to the distal femurs: skull base, spine (cervical, thoracic, lumbar), pelvis, and femurs. Ten slices were randomly selected from each segment (Python random module, version 3.8). As such, 60 slices were included for each subject in the dataset. The outline of the osseous skeleton on each slice was manually outlined by a medical student (MHC) using rilcontour [20], and these were reviewed by a musculoskeletal radiologist (FIB) (Fig. 2). The annotated image slices were then converted to 512×512 pixels Portable Network Graphics (PNG) for training a 2D U-net model.

Data preprocessing and annotation for lesion detection

CT images for object detection modeling were converted from DICOM files to PNG format. Subsequently, consecutive image slices from the skull base to the mid-femur were presented to two fellowship-trained musculoskeletal radiologists (FIB and MDR) with 2 and 11 years of post-training subspecialty expertise in interpreting whole-body low-dose CTs. The radiologists independently identified lytic lesions characteristic of myeloma on the image slices with bounding boxes using Labelme [21] (Fig. 3).

Model development for skeleton segmentation

A 2D Unet model from the MONAI framework was trained for the skeleton segmentation task with DiceLoss as the loss function and Adam as the optimizer [22] (Fig. 4a). Crop by positive and negative labels, rotation, flip, zooming, and affine were all random to ensure the model was generalizable and to avoid overfitting. The model was trained for 100 epochs and weighted based on the highest performance on

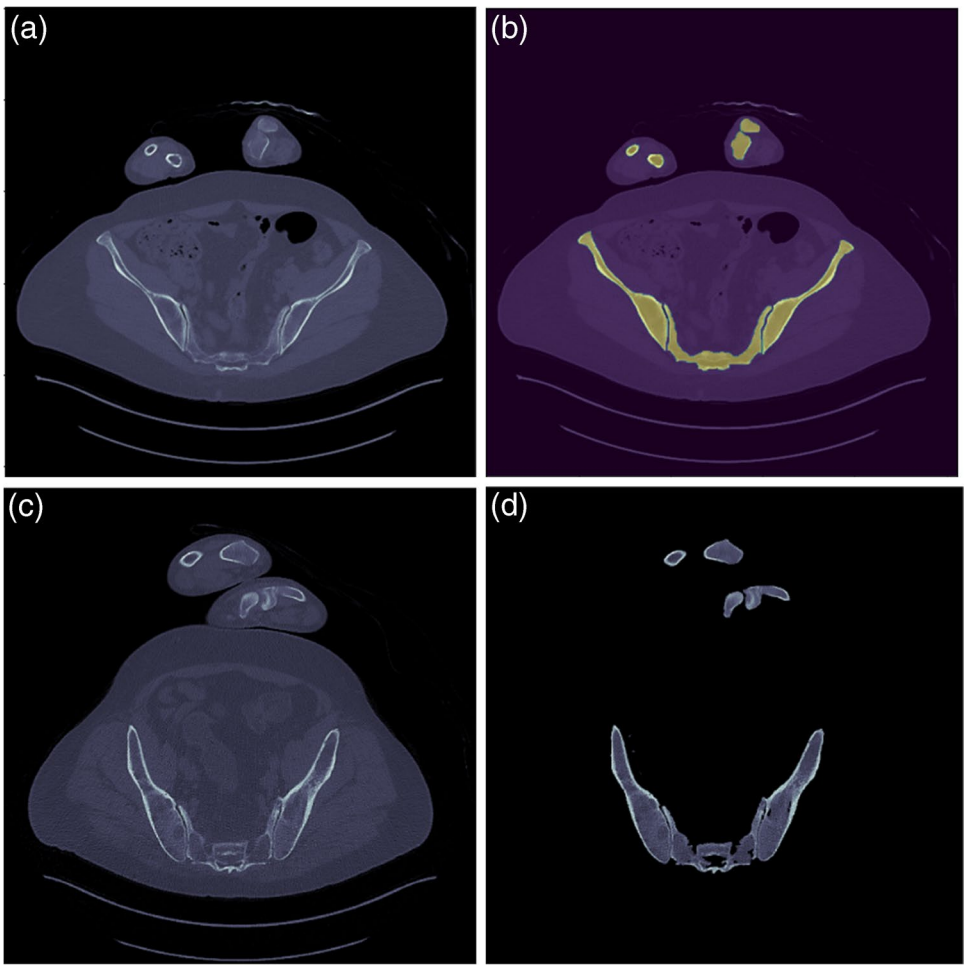

Fig. 2 Illustration of an example of manual bone annotations on an axial CT image, U-net prediction followed by Hadamard prediction. One sample of training set without annotation (**a**). One sample of training set with overlaid manual annotation (**b**) One sample of the test set (**c**). One sample of test set model's prediction followed by a Hadamard production with the original image (**d**)

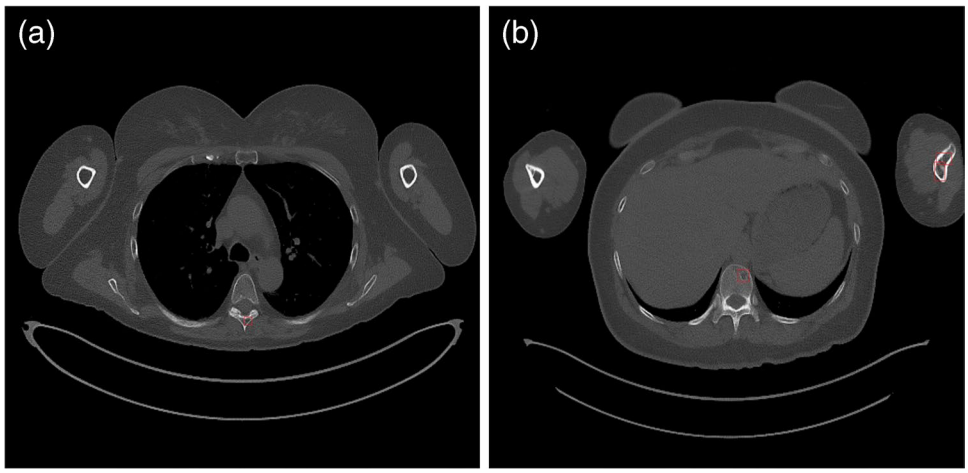

Fig. 3 Illustration of multiple myeloma lesion annotation. Red bounding boxes denote radiologist segmented lesions in the laminar of a thoracic vertebral body (**a**), a lesion in a thoracic vertebral body (**b**) and a lesion in the left humerus (**b**)

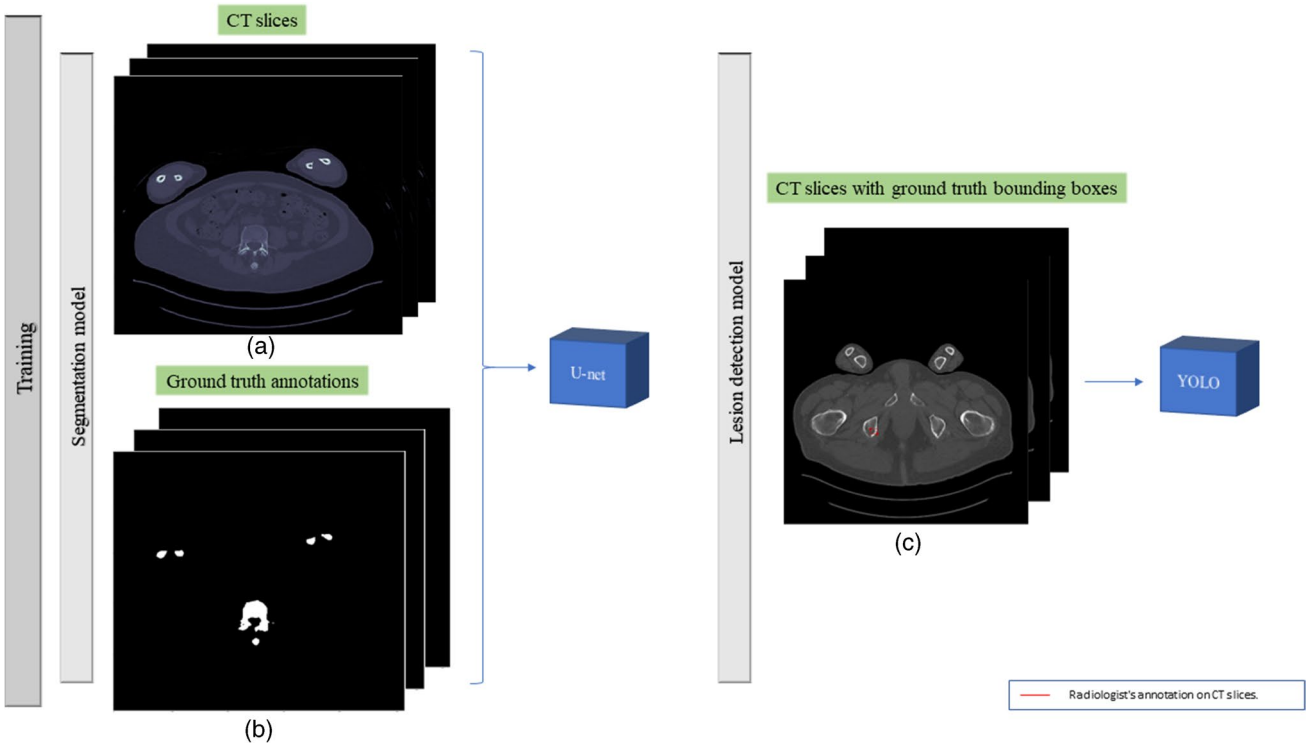

Fig. 4 Preprocessing and training pipeline of segmentation and lesion detection models

the validation set. Hyperparameters were optimized using the validation set (Supplementary Table 1).

Model development for lesion detection

"You Only Look Once" (YOLO) v5l was chosen as the model architecture due to its superior performance on the validation set and faster efficiency during inference compared to other object detection models [23, 24]. Since the ensemble theorem states predictions based on a combination of independent predictors can be more accurate than individual ones, we trained and optimized three different detection models with different data shuffling [25]. A variety of data augmentation techniques were utilized including horizontal and vertical flipping, scaling, translating, changing Hue, saturation, value (HSV), and mosaics for data augmentation (Fig. 4b). Further details of model development are provided in the supplemental Table 1.

Model hyperparameters were optimized using grid search (Supplemental Table 2). For inference on the test set, we run the ensemble of three models against all slices from the skull base to the distal femur. To determine the end-to-end performance of this two-step approach for

Table 1 Participant characteristics and plasma cell dyscrasia type. IQR, interquartile range

Subject characteristics	All subjects ($N=40$)
Median age in years (IQR)	67 (11)
Age range in years	35–90
Female	18/40 (45%)
Male	22/40 (55%)
Disease	
Multiple myeloma	28/40 (70%)
Smoldering multiple myeloma	1/40 (2.5%)
Myeloma with primary amyloidosis	2/40 (5%)
Monoclonal gammopathy of undetermined significance	6/40 (15%)
POEMS syndrome	2/40 (5%)
Solitary plasmacytoma	1/40 (2.5%)
Test cases	
Multiple myeloma with lytic bone lesions	3/4 (75%)
POEMS Syndrome with no lytic lesion	1/4 (25%)

detecting lytic lesions, 30 axial slices were randomly selected from each participant in the testing set and analyzed by the algorithm. The detected lesions were then reviewed in consensus by the MSK radiologists (Fig. 5).

All image processing and model development was performed using Pytorch 1.8.1, and MONAI 0.6 on Python 3.8 on a GPU cluster of 4 GPUs (NVIDIA A100).

Statistical methods

To determine the performance of each model, the Dice similarity coefficient (DSC) and mean average precision (mAP) at a 0.5 threshold were calculated. Receiver operating characteristic curve analysis was used to compare the sensitivity and specificity of the final model in detecting lytic lesions of multiple myeloma. Using different confidence thresholds for the YOLO detection model, we calculated the per lesion sensitivity and specificity, and the diagnostic performance was determined using the area under the receiver operating characteristic curve (AUROC). All statistical analyses were performed using scikit-learn 1.1 [26] on Python version 3.8.

Results

Forty unique subjects with biochemically confirmed plasma cell dyscrasias were included in the study. Demographics are summarized in Table 1. The median age of subjects was 67 (interquartile range = 11).

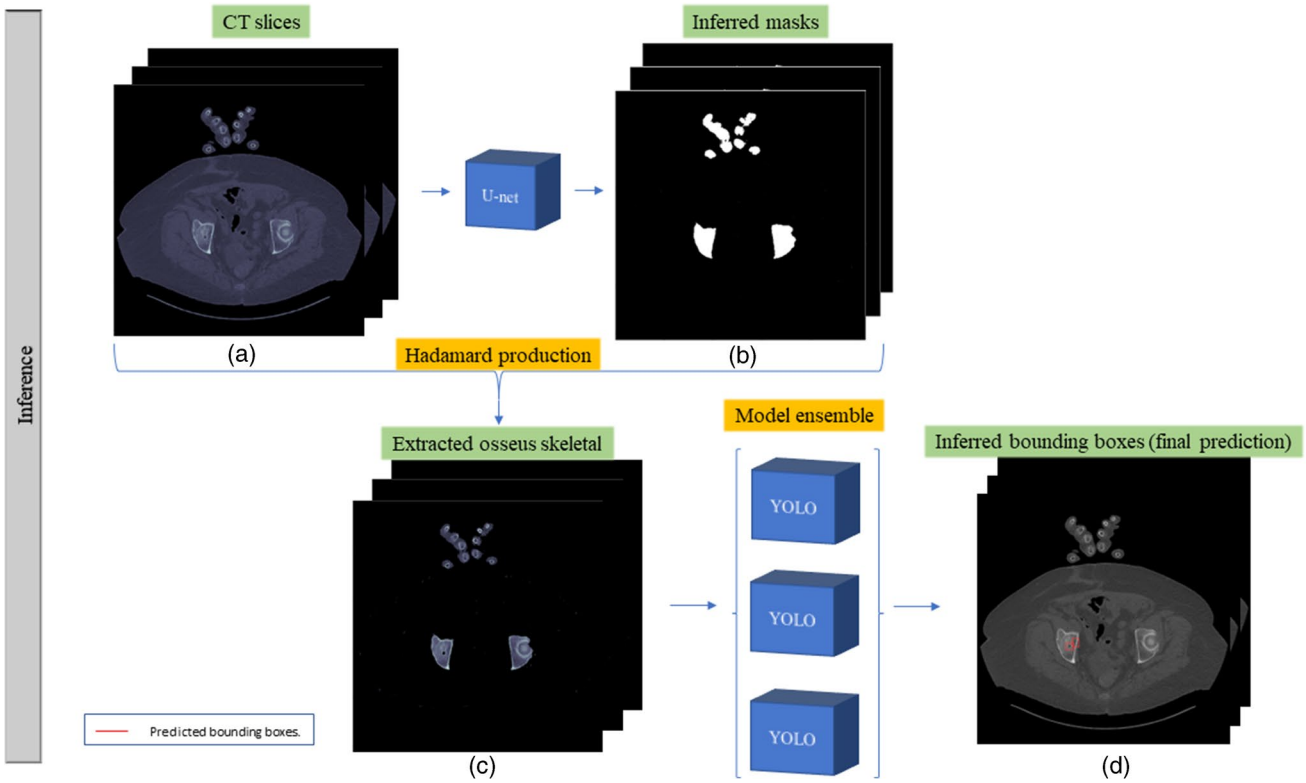

Fig. 5 Illustration of the two-step end-to-end multiple myeloma lesion detection model

Interobserver agreement

Following an independent review of 776 image slices from one test patient with biochemically confirmed multiple myeloma, the two radiologists (FIB and MDR) identified lytic lesions of multiple myeloma on 10 and 9 image slices, respectively. Interreader agreement was excellent (kappa = 0.84; 95% confidence interval: 0.66 to 1.00).

Model performance

Five hundred forty axial images were annotated for the bone segmentation model development. In total, 5640 lytic lesions were identified on 2193 axial slices (Fig. 1). The segmentation model achieved DSC of 0.93 (standard deviation (SD) = 0.03) and the lesion detection model achieved mAP.0.5: 0.47 (SD = 0.02). The final myeloma detection model achieved a sensitivity and specificity of 91.6% and 84.6%, respectively, at detecting lytic lesions. The model performance, as determined by the AUROC, was 90.4% (Fig. 6).

Discussion

Lytic bone lesions identified on whole-body low-dose skeletal surveys are a component of the imaging diagnosis of multiple myeloma and for therapy response assessment. In this study, we sought to develop a DL algorithm capable of detecting lytic bone lesions on a whole-body low-dose CT dataset. Using a two-step approach, consisting of osseous segmentation followed by lesion detection, we developed an end-to-end algorithm, which when applied to de novo axial whole-body low-dose CT slices, the model achieved a sensitivity and specificity of 91.6% and 84.6%, respectively, with an AUROC of 90.4% at detecting lytic lesions. CT skeletal surveys are superior to the conventional radiographic skeletal survey for osteolytic disease, yet the sheer volume of image slices substantially increases the length of time required to interpret these examinations. The end-to-end algorithm described in this study can augment the interpreter's effort by detecting lytic lesions.

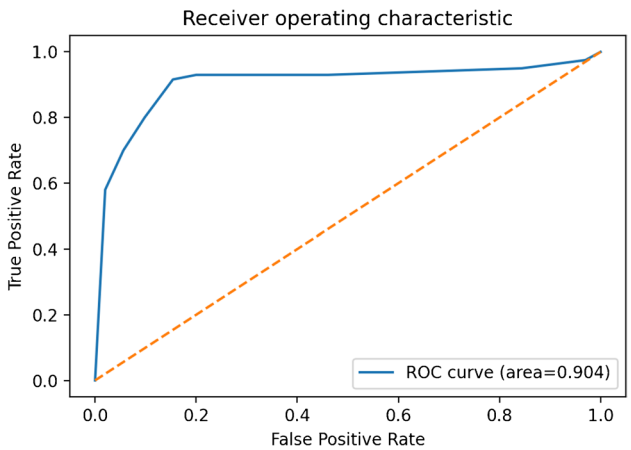

Fig. 6 Area under the receiver operating characteristic curve

Common object detection tasks in musculoskeletal imaging have reported accuracies ranging between 73% and 96% [27]. For instance, a model for detecting sclerotic spinal lesions had a sensitivity of 92% and specificity of 87% [28]. A two-step DL algorithm for detecting subtle fractures on radiographs had an accuracy of 90.8% for triquetral fractures and 92.5% for Segond fractures [18]. Another DL model to detect acute appendicular fractures in pediatric patients had a sensitivity and specificity of 91.3% and 90%, respectively [29]. While most of these detection algorithms focused on specific anatomic regions, the algorithm we developed in this study detects lytic lesions from the skull to the mid-femur.

Machine learning has been applied to the biochemical diagnosis and classification of multiple myeloma, including bone marrow aspirate and flow cytometric analysis [15, 30]. Radiologic evaluation and machine learning algorithms for lesion detection have been tested on positron emission tomography/CT [14, 15] and to determine the radiomic features that distinguish myeloma lesions from metastatic lesions on lumbar spine MRIs [31]. Whole-body low-dose CT is the recommended initial imaging modality for multiple myeloma, and as such, the DL algorithm presented in this study is a valuable tool for both the early imaging diagnosis of multiple myeloma and assessment of disease progression.

Despite the unique imaging features of lytic bone lesions, mimickers such as rounded intramedullary foci of osteopenia resulted in false-positive detections (Fig. 7). On the other hand, an expansile lytic lesion with destruction of the overlying cortex (Fig. 8) was not detected by the algorithm (false negative). Although an extra-osseous extension of multiple myeloma is common, the training data did not include such lesions with extensively disrupted cortex.

This study has a few limitations. First, although we used a very large number of slices, there was a relatively small number of subjects in this study. Second, all datasets used for algorithm training and testing were acquired on CT systems from one vendor at a single institution. Although the CTs were obtained from two scanner models, all images were acquired with the same protocol and section thickness. It would be important to validate this algorithm on datasets acquired on CT systems from other vendors and from various institutions with different imaging protocols.

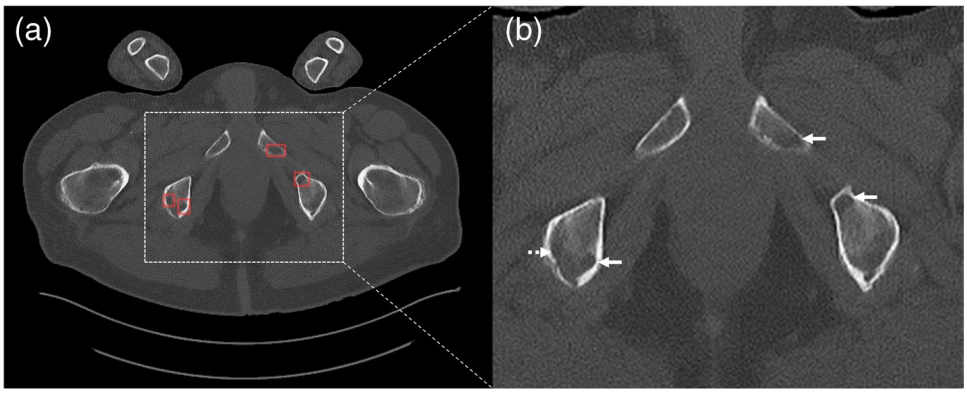

Fig. 7 Seventy-three-year-old male with multiple myeloma. Axial image through the ischial tuberosities (**a**) and cut-out magnified image (**b**). Red boxes (**a**) outline the lesions detected by the algorithm. Solid white arrows denote the true lytic lesions (confirmed with multislice evaluation of the region). The dashed white arrow denotes a region of focal osteopenia incorrectly detected as a lytic lesion (false positive)

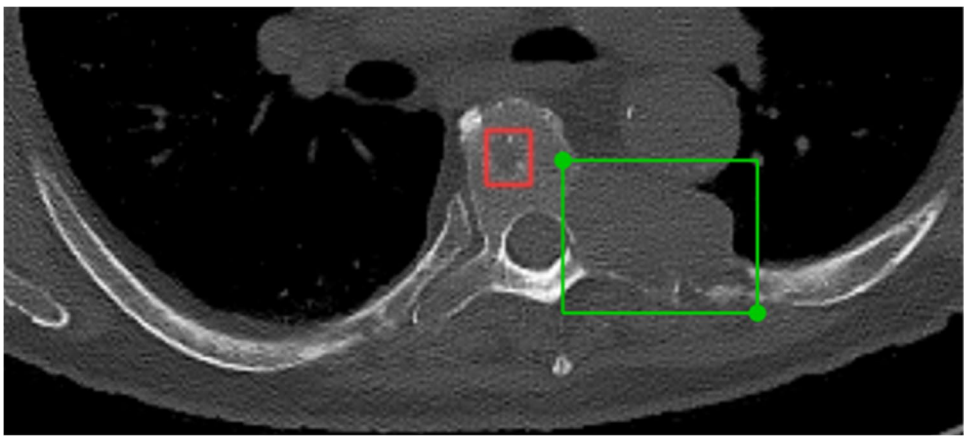

Fig. 8 Seventy-eight-year-old male with multiple myeloma. Axial image through the thoracic spine. The red box outlines a lytic lesion detected by the algorithm. This was confirmed as a true lesion after a multislice image review. The green box outlines an expansile lytic lesion involving the left posterior elements of the thoracic vertebral body and the adjacent rib with the destruction of the overlying cortex. This was not detected by the algorithm (false negative)

In conclusion, we demonstrate the development and internal validation of a novel, two-step DL model to identify multiple myeloma lesions on whole-body low-dose CT images. The first part of this two-step algorithm—bone segmentation—defines a region of interest on which the second part of the algorithm—lesion detection—is applied. This approach reduces the background noise effect from the soft tissues, potentially obviating the need for massive training datasets. Future directions for this study include external validation of this algorithm and quantification of osteolytic disease burden in patients with multiple myeloma.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00256-022-04160-z.

Declarations

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Conflicts of interest Bradley Erickson: officer: FlowSIGMA, Inc. and Yunu, Inc. The other others declare no competing interests.

References

1. Durie BG, Salmon SE. A clinical staging system for multiple myeloma. Correlation of measured myeloma cell mass with presenting clinical features, response to treatment, and survival. Cancer. 1975;36:842–54.
2. Hillengass J, Usmani S, Rajkumar SV, Durie BGM, Mateos M-V, Lonial S, et al. International myeloma working group consensus recommendations on imaging in monoclonal plasma cell disorders. Lancet Oncol. 2019;20:e302–12.
3. Moulopoulos LA, Koutoulidis V, Hillengass J, Zamagni E, Aquerreta JD, Roche CL, et al. Recommendations for acquisition, interpretation and reporting of whole body low dose CT in patients with multiple myeloma and other plasma cell disorders: a report of the IMWG Bone Working Group. Blood Cancer J. 2018;8:95.
4. Horger M, Claussen CD, Bross-Bach U, Vonthein R, Trabold T, Heuschmid M, et al. Whole-body low-dose multidetector row-CT in the diagnosis of multiple myeloma: an alternative to conventional radiography. Eur J Radiol. 2005;54:289–97.
5. Gavriatopoulou M, Boultadaki A, Koutoulidis V, Ntanasis-Stathopoulos I, Bourgioti C, Malandrakis P, et al. The role of low dose whole body CT in the detection of progression of patients with smoldering multiple myeloma. Blood Cancer J. 2020;10:93.
6. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: state of the art and future directions. J Digit Imaging. 2017;30:449–59.
7. Weston AD, Korfiatis P, Kline TL, Philbrick KA, Kostandy P, Sakinis T, et al. Automated abdominal segmentation of CT scans

for body composition analysis using deep learning. Radiology. 2019;290:669–79.
8. Kline TL, Korfiatis P, Edwards ME, Blais JD, Czerwiec FS, Harris PC, et al. Performance of an artificial multi-observer deep neural network for fully automated segmentation of polycystic kidneys. J Digit Imaging. 2017;30:442–8.
9. Korfiatis P, Kline TL, Erickson BJ. Automated segmentation of hyperintense regions in FLAIR MRI using deep learning. Tomography. 2016;2:334–40.
10. Qu R, Yang Y, Wang Y. COVID-19 detection using CT image based on YOLOv5 network [Internet]. arXiv [eess.IV]. 2022. Available from: http://arxiv.org/abs/2201.09972
11. Hossain A, Islam MT, Almutairi AF. A deep learning model to classify and detect brain abnormalities in portable microwave based imaging system. Sci Rep. 2022;12:6319.
12. Vyshnav MT, Sowmya V, Gopalakrishnan EA, Variyar V.V. S, Menon VK, Soman P K. Deep learning based approach for multiple myeloma detection. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2020. p. 1–7.
13. He J, Zhang K. Medical image analysis of multiple myeloma based on convolutional neural network. Expert Syst [Internet]. Wiley; 2022;39. Available from: https://onlinelibrary.wiley.com/doi/https://doi.org/10.1111/exsy.12810
14. Xu L, Tetteh G, Lipkova J, Zhao Y, Li H, Christ P, et al. Automated whole-body bone lesion detection for multiple myeloma on 68Ga-pentixafor PET/CT imaging using deep learning methods. Contrast Media Mol Imaging. 2018;2018:2391925.
15. Allegra A, Tonacci A, Sciaccotta R, Genovese S, Musolino C, Pioggia G, et al. Machine learning and deep learning applications in multiple myeloma diagnosis, prognosis, and treatment selection. Cancers [Internet]. 2022;14. Available from: http://dx.doi.org/https://doi.org/10.3390/cancers14030606
16. Faghani S, Codipilly DC, Vogelsang D, Moassefi M, Rouzrokh P, Khosravi B, et al. Development of a deep learning model for the histological diagnosis of dysplasia in Barrett's esophagus. Gastrointest Endosc [Internet]. 2022; Available from: https://www.sciencedirect.com/science/article/pii/S0016510722017643
17. Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: a review [Internet]. arXiv [cs.LG]. 2021. Available from: http://arxiv.org/abs/2104.02395
18. Ren M, Yi PH. Deep learning detection of subtle fractures using staged algorithms to mimic radiologist search pattern. Skeletal Radiol. 2022;51:345–53.
19. Franklin J. The elements of statistical learning: data mining, inference and prediction. Math Intelligencer. 2005;27:83–5.
20. Philbrick KA, Weston AD, Akkus Z, Kline TL, Korfiatis P, Sakinis T, et al. RIL-Contour: a medical imaging dataset annotation tool for and with deep learning. J Digit Imaging. 2019;32:571–81.
21. Russell BC, Torralba A, Murphy KP, Freeman WT. LabelMe: a database and web-based tool for image annotation. Int J Comput Vis. 2008;77:157–73.
22. MONAI Consortium. MONAI: medical open network for AI [Internet]. 2022. Available from: https://zenodo.org/record/6639453
23. Girshick R, Donahue J, Darrell T, Malik J. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition. p. 580–7.
24. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection [Internet]. arXiv [cs.CV]. 2015. Available from: http://arxiv.org/abs/1506.02640
25. Hansen LK, Salamon P. Neural network ensembles. IEEE Trans Pattern Anal Mach Intell. 1990;12:993–1001.
26. Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, et al. Scikit-learn: machine learning in Python. J Mach Learn Res [Internet]. Available from: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html
27. Yang S, Yin B, Cao W, Feng C, Fan G, He S. Diagnostic accuracy of deep learning in orthopaedic fractures: a systematic review and meta-analysis. Clin Radiol. 2020;75:713.e17-713.e28.
28. Chang CY, Buckless C, Yeh KJ, Torriani M. Automated detection and segmentation of sclerotic spinal lesions on body CTs using a deep convolutional neural network. Skeletal Radiol. 2022;51:391–9.
29. Hayashi D, Kompel AJ, Ventre J, Ducarouge A, Nguyen T, Regnard N-E, et al. Automated detection of acute appendicular skeletal fractures in pediatric patients using deep learning. Skeletal Radiol [Internet]. 2022; Available from: http://dx.doi.org/https://doi.org/10.1007/s00256-022-04070-0
30. Yan W, Shi H, He T, Chen J, Wang C, Liao A, et al. Employment of artificial intelligence based on routine laboratory results for the early diagnosis of multiple myeloma. Front Oncol. 2021;11:608191.
31. Xiong X, Wang J, Hu S, Dai Y, Zhang Y, Hu C. Differentiating between multiple myeloma and metastasis subtypes of lumbar vertebra lesions using machine learning-based radiomics. Front Oncol. 2021;11:601699.
32. Lin T-Y, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, et al. Microsoft COCO: common objects in context [Internet]. arXiv [cs.CV]. 2014. Available from: http://arxiv.org/abs/1405.0312